# 컨텍스트 반영 제스처 이해를 통한 아이템 선택: 실험적 검증과 프레임워크 제안*

박예서[0], 오승재
경희대학교 소프트웨어융합학과
yeseo@khu.ac.kr, oreo329@khu.ac.kr

# Item Selection through Context-Aware Gesture Understanding: Experimental Validation and Framework Proposal

Yeseo Park[0], Seungjae Oh
Department of Software Convergence, Kyung Hee University

**요 약**

가상현실과 증강현실 환경에서 제스처 인식은 직관적인 상호작용을 가능하게 하는 핵심 기술로 주목받아 왔다. 그러나 기존 연구들은 주로 인식 정확도에 집중하여 실제 맥락을 반영한 의도 파악에는 한계가 있었다. 본 연구에서는 컨텍스트를 상황 요소(전투, 요리, 제작)로 정의하고, 이를 활용하여 사용자가 인벤토리 내 특정 아이템을 사용하는 제스처를 취했을 때, 보다 의도에 부합하는 아이템 자동 선택이 가능함을 검증한다. MediaPipe로 수집한 제스처 데이터를 활용해 동일한 모델 기반에서 컨텍스트 입력의 유무를 비교한 결과, 컨텍스트를 반영했을 때 아이템 선택의 정확도가 뚜렷하게 증가하였다. 이를 통해 컨텍스트가 제스처 기반 인터랙션의 정확성과 일관성에 기여함을 확인하였으며, 나아가 제스처 인코더와 컨텍스트 인코더를 결합한 소프트트리 기반 프레임워크를 제안한다.

## 1. 서 론

가상현실과 증강현실 환경에서 제스처 인식은 직관적이고 자연스러운 상호작용을 가능하게 한다. 이는 인간의 신체 움직임을 직접 입력으로 활용하는 대표적인 인간-컴퓨터 상호작용(HCI) 방식으로, 별도의 장치 없이도 사용자가 시스템과 자연스럽게 소통할 수 있다는 점에서 학문적 및 실용적 가치가 크다. 특히 손 자체를 인터페이스로 활용하는 연구는 손동작으로 도구와 객체를 모사하여 시스템의 대응 항목과 연계함으로써 표현 범위를 넓히고 상호작용의 자연스러움을 높인 바 있다 [1]. 반면 기존의 많은 접근은 제스처의 패턴 분류 정확도 향상에 초점을 두었고, 시간 정보를 효율적으로 결합해 성능을 개선하는 것과 같은 방법들이 보고되었다 [2]. 그러나 이러한 흐름은 장면과 활동 같은 맥락을 충분히 반영하지 못할 때 사용자의 실제 의도가 결과에 온전히 반영되기 어렵다는 한계를 드러낸다.

이 한계를 극복하기 위해 다양한 맥락 정보와 제스처 데이터를 결합하는 연구가 등장하고 있다. 로봇 제어 분야에서는 제스처를 개별 동작 단위가 아닌 에피소드로 정의하고 상황 정보를 함께 반영함으로써 사용자의 목표 의도를 더 정확히 추론하는 방법이 제안되었다 [3]. 또한 최근에는 대규모 언어 모델을 활용하여 자유형 제스처의 의미를 추론하고, 시선이나 환경 상태와 같은 주변 맥락을 대화적으로 통합하여 최종 명령으로 연결하는 프레임워크가 보고되었다 [4][5]. 이러한 접근은 제스처 인식이 단순히 패턴 매칭을 넘어, 다양한 맥락적 요소와 결합할 때 의미 해석의 정확성이 높아질 수 있음을 보여준다.

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2025년도 SW중심대학사업의 결과로 수행되었음(2023-0-00042)

제스처의 의미를 언어적 지식과 연결하여 인식 성능을 강화하려는 시도도 활발히 진행되고 있다. 스켈레톤 기반 동작 인식 연구에서는 동작의 언어적 설명을 자동으로 생성하여 이를 학습 과정에 통합함으로써, 동작과 신체 부위 간의 관계를 사전 지식 형태로 모델에 주입하는 방법이 제안되었다 [6]. 이와 같은 다중 모달 학습은 기존 방식 대비 성능 향상을 가져왔으며, 제스처와 의미적 표현을 연결하는 가능성을 제시하였다. 특히 이러한 정렬 과정은 임베딩 공간에서 제스처와 의미적 단서 간의 유사도를 학습하는 대조 손실 기법과 맞닿아 있으며, 새로운 객체나 상황에서도 유연하게 확장될 수 있는 기반을 마련한다는 점에서 주목할 만하다.

이러한 연구 흐름은 제스처 인식에서 맥락을 통합하는 것이 단순한 정확도 향상을 넘어 사용자의 의도를 해석하는 핵심 요소임을 보여준다. 본 연구에서는 이를 확인하기 위해 상황 정보를 컨텍스트로 활용한 소규모 실험을 수행하여 맥락이 아이템 선택에 미치는 효과를 검증하였다. 나아가 제스처와 언어적 표현을 연결해 의미적 단서를 강화하고, 상황 외에도 다양한 컨텍스트 요소를 결합할 수 있는 확장 가능성을 고려하여, 제스처 인코더와 컨텍스트 인코더를 통합한 소프트트리 기반 프레임워크를 제안한다.

## 2. 실험 설계 및 방법

### 2.1 실험 개요

본 연구는 컨텍스트를 활용할 때 제스처 기반 아이템 선택의 정확도가 얼마나 향상되는지를 검증하는 데 목적이 있다. 이를 위해 전투, 요리, 제작의 세 가지 상황을 가정하고, 동일한 제스처 데이터에 대해 컨텍스트 입력의 유무를 비교하였다. 모델은

TD-GCN 기반 제스처 인코더 [7]의 임베딩을 백본으로 활용하여 동일한 조건에서 학습되도록 설계하였다. 이를 통해 모델 구조나 파라미터 차이가 아닌, 컨텍스트 정보 자체가 성능 향상에 기여하는 정도를 직접적으로 평가할 수 있도록 하였다.

### 2.2 데이터 수집

제스처 데이터는 MediaPipe Hands를 이용하여 21개 손 관절 좌표 시퀀스로 추출하였다. 2명의 참가자가 인벤토리 내 특정 아이템을 실제로 사용하는 상황을 상상하며 2초간 손동작을 취하였다. 전투 상황에서는 검과 권총, 요리 상황에서는 식칼과 프라이팬, 제작 상황에서는 망치와 톱에 대해 수행하였다. 각 아이템 제스처는 참가자당 12회 반복 촬영되었으며, 최초 프레임의 손목 위치를 기준으로 모든 관절 좌표를 상대 좌표로 변환하고, 손 크기에 따른 차이를 줄이기 위해 거리 기반 스케일 정규화를 수행하였다. 또한, 각 시퀀스에는 전투, 요리, 제작의 세 상황을 확률로 표현한 3차원 컨텍스트 벡터가 함께 생성되었다. 해당 시퀀스의 실제 상황에 대응하는 확률이 0.5 이상이 되도록 설정하였으며, 세 확률의 합은 항상 1이 되도록 랜덤 생성하였다.

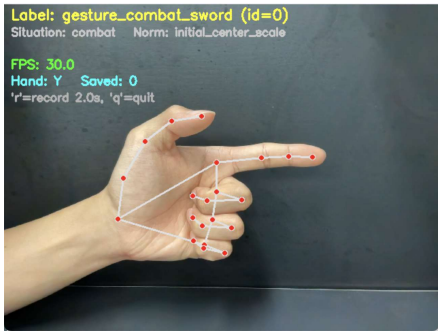


그림 1. MediaPipe을 통한 시퀀스 수집

### 2.3 입력 표현 및 모델 구조

입력 데이터는 제스처 시퀀스와 컨텍스트 벡터(context prior)로 구성된다. 제스처 시퀀스는 TD-GCN 기반의 제스처 인코더에 입력된다. 모델은 공개된 제스처 인코더의 구조를 그대로 활용하되, 최종 14차원 분류 헤더를 제거하고 그 이전의 임베딩 벡터를 백본으로 사용하였다. 이 임베딩은 아이템 선택을 위한 로짓(logit)을 계산하는 데 사용되며 컨텍스트 벡터는 가중 계수 $\alpha$와 함께 로짓에 더해져 제스처 임베딩의 의미적 표현을 강화한다.

### 2.4 학습 및 평가

모델 학습은 제스처 시퀀스와 컨텍스트 벡터를 입력으로 하여 아이템의 정답 레이블을 예측하도록 설계하였다. 손실 함수는 교차 엔트로피를 사용하였으며, 모든 조건에서 동일한 학습률과 최적화 알고리즘을 적용하였다. 컨텍스트 조건에서는 경험적으로 가중 계수 $\alpha$를 0.9로 설정하여 context prior가 로짓 계산에 반영되도록 하였고, 무컨텍스트 조건에서는 $\alpha$를 0으로 하여 동일한 구조에서 컨텍스트의 영향을 배제하였다. 데이터셋의 크기가 제한적이므로 5-fold 교차검증을 사용하였다. 각 fold에서 모델을 독립적으로 학습하고 평가하였으며, 예측 결과와 정답을 csv 형식으로 저장하였다. 평가 지표는 최종 선택된 아이템이 사용자의 의도와 일치하는 비율인 정확도를 사용하였다.

## 3. 실험 결과

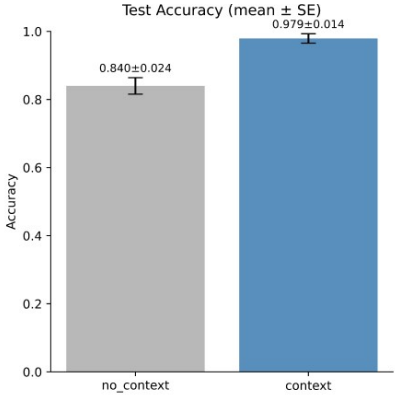


그림 2. 컨텍스트 유무에 따른 정확도 막대 그래프

### 3.1 컨텍스트 유무에 따른 정확도 비교

그림 2는 컨텍스트 입력의 유무에 따라 아이템 선택 정확도를 비교한 결과이다. 5-fold 교차검증 결과, 컨텍스트를 포함한 조건의 평균 정확도는 0.979(±0.014)로, 무컨텍스트 조건의 0.840(±0.024)보다 높게 나타났다. 또한 표준오차가 더 작아 컨텍스트를 반영한 모델이 전반적으로 안정적인 예측 성능을 보였다. 이 결과는 컨텍스트 정보가 제스처 표현의 의미적 구분을 보완하여 아이템 선택의 정확도와 일관성을 함께 향상시킨다는 것을 보여준다.

### 3.2 비모수적 검정

컨텍스트 입력이 모델의 성능에 미치는 영향을 통계적으로 검정하기 위해 5-fold 정확도 결과를 대상으로 Wilcoxon signed-rank 검정(one-sided) [8]을 수행하였다. 검정 결과, 통계량 W는 15.0, p값은 0.03125로 나타나 유의수준 0.05에서 컨텍스트를 포함한 조건의 정확도가 무컨텍스트 조건보다 유의하게 높다는 대립가설을 지지한다. Fold 간 정확도 차이의 중앙값은 +0.1379로, 모든 fold에서 컨텍스트를 포함한 조건이 더 높은 정확도를 보였다. 이는 컨텍스트 정보가 모델 전반에서 일관된 성능 향상을 가져왔음을 의미한다.

## 4. 제안하는 프레임워크

실험 결과를 통해 컨텍스트 정보가 제스처 기반 아이템 선택의 정확도를 높이는 데 기여함을 확인하였다. 이를 확장하여, 본 연구는 제스처 인코더와 컨텍스트 인코더를 결합하고 소프트트리를 활용하는 프레임워크를 제안한다. 제스처 인코더는 양손의 스켈레톤 좌표 벡터들의 시퀀스를 입력으로 받고, 컨텍스트 인코더는 주변 환경의 객체들과 그 관계들을 이해할 수 있도록 Scene Graph를 생성하여 그 시퀀스를 입력으로 받는다. 각 인코더의 출력이 결합된 임베딩은 LLM이 생성한 정답 아이템에 대해 affordance 및 맥락에 관한 텍스트들의 임베딩과 학습하며 의미적으로 정렬된다. 이렇게 정렬된 임베딩은 최종 아이템 선택에서 소프트트리의 확률적 분기를 통해 결정된다.

### 4.1 제스처 인코더

제스처 인코더 $E_g$는 프레임마다 추출된 양손 스켈레톤 좌표 벡터 시퀀스 $x_g$를 입력으로 제스처 임베딩 $z_g$을 생성한다.

$$z_g = E_g(x_g) \tag{1}$$

4.2 컨텍스트 인코더

컨텍스트 인코더 $E_c$는 같은 시간 동안 생성된 Scene Graph 시퀀스 벡터 $SG$를 입력으로 받아, 주변 환경에 대한 컨텍스트 임베딩 $z_c$를 생성한다.

$$z_c = E_c(SG, gaze) \quad (3)$$

여기서 gaze 정보도 입력하여, 사용자가 주목하는 영역 근처의 객체가 더 큰 가중치를 가지도록 반영한다.

4.3 정답 물체 임베딩

아이템 $i$는 텍스트 기반 affordance 및 어떤 맥락에서 사용될 수 있는지를 포함한 다양한 의미적 표현을 텍스트 인코더 $E_a$를 통해 정답 임베딩 $z_a(i)$로 표현한다.

$$z_a(i) = E_a(aff_i) \quad (2)$$

이를 결합된 임베딩과 대조 손실 [6]을 통해, 분류 전 더 다양한 표현과 그 의미를 정답에 가깝게 정렬할 수 있다.

4.4 소프트트리 기반 선택 모듈

최종 선택을 위해 먼저 제스처 임베딩 $z_g$와 컨텍스트 임베딩 $z_c$를 결합하여 정렬한 임베딩을 입력으로 게이트 확률 $p$를 산출한다.

$$p = \sigma(W_{gate}[z_g; z_c] + b_{gate}) \quad (4)$$

소프트트리의 두 리프는 아이템 유사도 점수를 출력한다.

$$y_l(i) = \tilde{z}_{g;c}^{\top} \tilde{z}_a(i) \quad (5)$$

최종 출력은 게이트 확률로 두 리프의 출력을 가중합하여 얻는다. 이때 가장 높은 확률을 갖는 아이템이 선택된다.

$$\hat{y}(i) = p \cdot y_{l1}(i) + (1-p) \cdot y_{l2}(i) \quad (6)$$

4.5 기대 효과

본 프레임워크의 첫 번째 특징은 컨텍스트가 융합된 제스처 표현과 아이템의 affordance 및 맥락적 표현을 같은 공간에서 직접 맞추도록 설계되었다는 점이다. 이를 통해 서로 다른 제스처라 하더라도 의미적으로 유사한 경우에는 가까운 위치로 모이고, 전혀 다른 의미를 가지는 경우에는 멀어지도록 학습된다. 이러한 구조는 단순히 제스처의 모양을 구분하는 수준을 넘어, 그 동작이 담고 있는 기능적 의미까지 반영할 수 있도록 한다.

두 번째 특징은 Scene Graph와 시선 정보를 함께 활용하는 컨텍스트 인코더이다. 이 인코더는 장면 내 객체 간의 관계를 분석하고, 사용자의 주의가 집중된 영역에 더 큰 비중을 두어 실제 상황의 맥락을 모델에 반영한다. 이를 통해 같은 제스처라 하더라도 상황에 따라 적절한 아이템이 달라지는 맥락적 변화를 효과적으로 반영할 수 있다.

마지막으로 소프트트리 구조는 제스처 임베딩과 컨텍스트 임베딩이 상호작용하는 방식을 확률적으로 모델링한다. 이는 특정 규칙에 따라 단순 분기하는 것이 아니라, 상황과 제스처의 조합에 따라 부드럽게 가중치를 조정하면서 아이템을 선택할 수 있게 한다. 결과적으로 새로운 아이템이 추가되더라도 기존 구조를 유지하면서 유연하게 확장할 수 있는 장점을 제공한다.

**5. 결론 및 향후 과제**

본 연구는 제스처 인식 과정에 상황 정보를 컨텍스트로 통합함으로써 인벤토리 내 아이템 선택의 정확성을 향상시킬 수 있음을 소규모 실험을 통해 확인하였다. 전투, 요리, 제작이라는 세 가지 상황을 설정하여 비교한 결과, 동일한 구조에서 컨텍스트를 포함했을 때 더 높은 정확도를 보였다. 이는 맥락적 단서가 제스처 기반 상호작용의 결과를 실제 의도와 일치하도록 조정하는 핵심 요소임을 보여준다.

이러한 결과를 바탕으로, 제스처 인코더와 컨텍스트 인코더를 결합한 소프트트리 기반 프레임워크를 제안하였다. 제스처 임베딩은 컨텍스트 임베딩과 융합되어 아이템의 affordance 및 맥락적 표현의 임베딩과 의미적으로 정렬되도록 학습되고, Scene Graph와 gaze 정보를 활용한 컨텍스트 인코더는 실제 장면의 맥락을 효과적으로 반영한다. 나아가 소프트트리 구조는 제스처와 컨텍스트 간의 상호작용을 확률적으로 모델링하여, 단순 분류기를 사용하는 방식보다 유연하고 확장 가능한 아이템 선택을 가능하게 한다.

향후 연구에서는 제스처와 상황을 보다 다양한 시나리오로 확장하여 일반화 가능성을 검증하는 것이 필요하다. 또한 Scene Graph 기반의 장면 정보뿐만 아니라 사용자 상태나 환경적 제약과 같은 복합적인 맥락을 함께 반영할 수 있는 인코더 구조로 확장할 수 있다. 더 나아가 실제 가상현실 및 증강현실 환경에서 실시간으로 동작하는 프로토타입을 구현하고 사용자 연구를 수행함으로써, 제안된 프레임워크가 상호작용 경험 전반에 미치는 효과를 검증하는 과정이 요구된다.

결과적으로 본 연구에서는 제스처 인식이 단순한 패턴 분류를 넘어, 사용자의 의도와 맥락을 함께 고려하는 방향으로 나아가야 한다는 점을 강조하며, 수행한 실험과 제안한 프레임워크는 이러한 전환을 실현하기 위한 기초적 시도라는 점에서 의의를 가진다.

**참 고 문 헌**


[1] Pei, Siyou, et al. "Hand interfaces: Using hands to imitate objects in ar/vr for expressive interactions." Proceedings of the 2022 CHI conference on human factors in computing systems. 2022.

[2] Lin, Ji, Chuang Gan, and Song Han. "Tsm: Temporal shift module for efficient video understanding." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[3] P. Vanc, J. K. Behrens and K. Stepanova, "Context-aware robot control using gesture episodes," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, pp. 9530-9536, 2023.

[4] Zeng, Xin, et al. "GestureGPT: Toward zero-shot free-form hand gesture understanding with large language model agents." Proceedings of the ACM on Human-Computer Interaction 8.ISS, pp. 462-499, 2024.

[5] Kobzarev, Oleg, Artem Lykov, and Dzmitry Tsetserukou. "Gestllm: Advanced hand gesture interpretation via large language models for human-robot interaction." 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 2025.

[6] Xiang, Wangmeng, et al. "Generative action description prompts for skeleton-based action recognition." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

[7] J. Liu, X. Wang, C. Wang, Y. Gao and M. Liu, "Temporal Decoupling Graph Convolutional Network for Skeleton-Based Gesture Recognition," in IEEE Transactions on Multimedia, vol. 26, pp. 811-823, 2024.

[8] Demšar, Janez. "Statistical comparisons of classifiers over multiple data sets." Journal of Machine learning research 7.Jan, pp.1-30, 2006.